Report: Clustering for Customer Segmentation

1. Main Objective of the Analysis

The main objective of this analysis is to perform customer segmentation using clustering techniques. The focus will be on identifying distinct customer groups based on purchasing behavior and other attributes. This analysis will benefit the business by providing insights into customer profiles, enabling targeted marketing strategies, improving customer service, and increasing customer retention.

2. Brief Description of the Dataset

I employed a UK-based online retail dataset obtained from the UCI machine learning repository https://archive.ics.uci.edu/dataset/352/online+retail. The retail dataset consists of 541,909 customer records and eight features:

- InvoiceNo : a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
- StockCode: a 5-digit integral number uniquely assigned to each distinct product.
- Description: product name
- Quantity: the quantities of each product (item) per transaction.
- InvoiceDate: the day and time when each transaction was generated.
- UnitPrice: product price per unit.
- CustomerID: a 5-digit integral number uniquely assigned to each customer.
- Country: the name of the country where each customer resides.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	United Kingdom

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	United Kingdom

3. Data Exploration and Cleaning

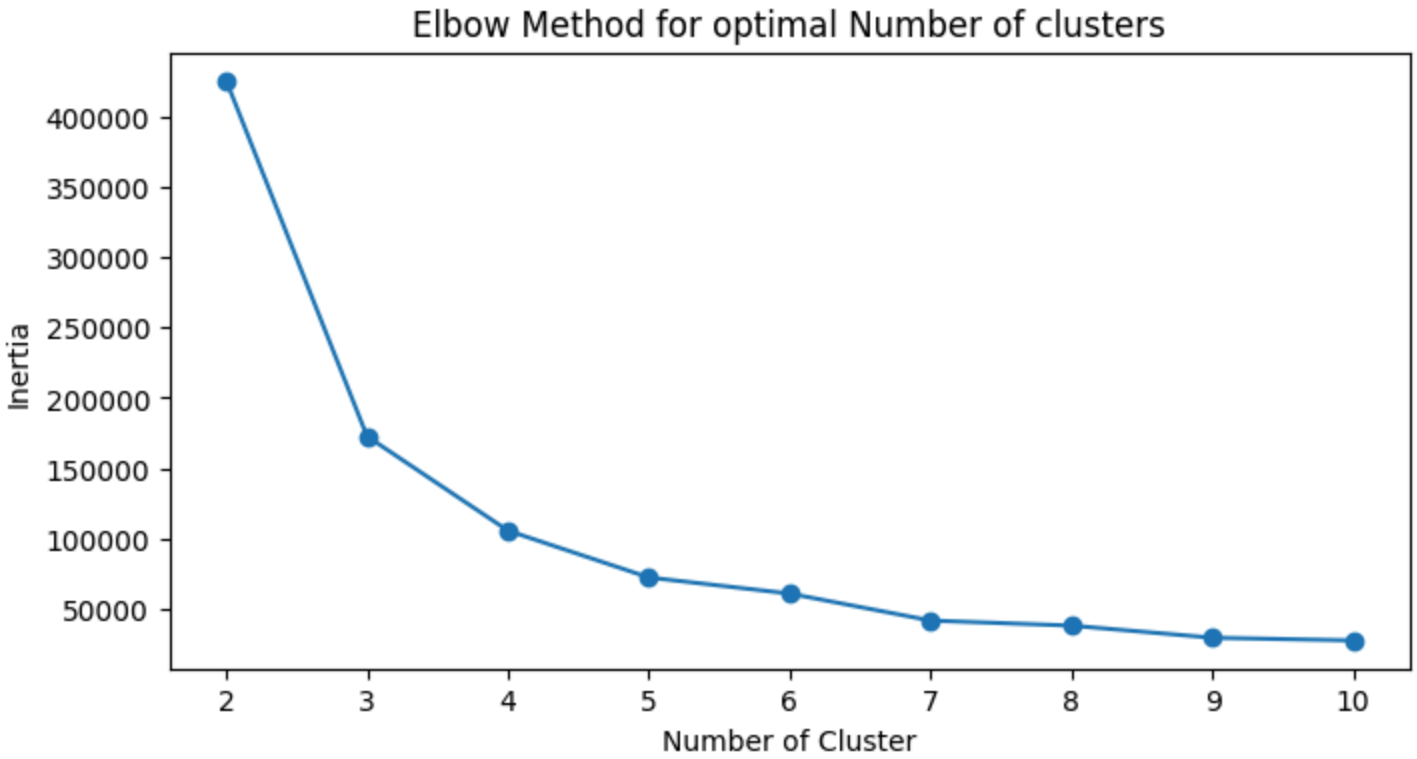
Data Pre-Processing:

- During the pre-processing stage, the null values were identified, specifically in the "Description" and "CustomerID" fields. By successfully remvoing null values for "CustomerID" entries.
- Removed 8905 incorrect values from the column "quantity" in the dataset.
- Also removed duplicate records by removing rows with identical values in all columns.

4. Training Variations of Unsupervised Models

Model 1: K-Means Clustering

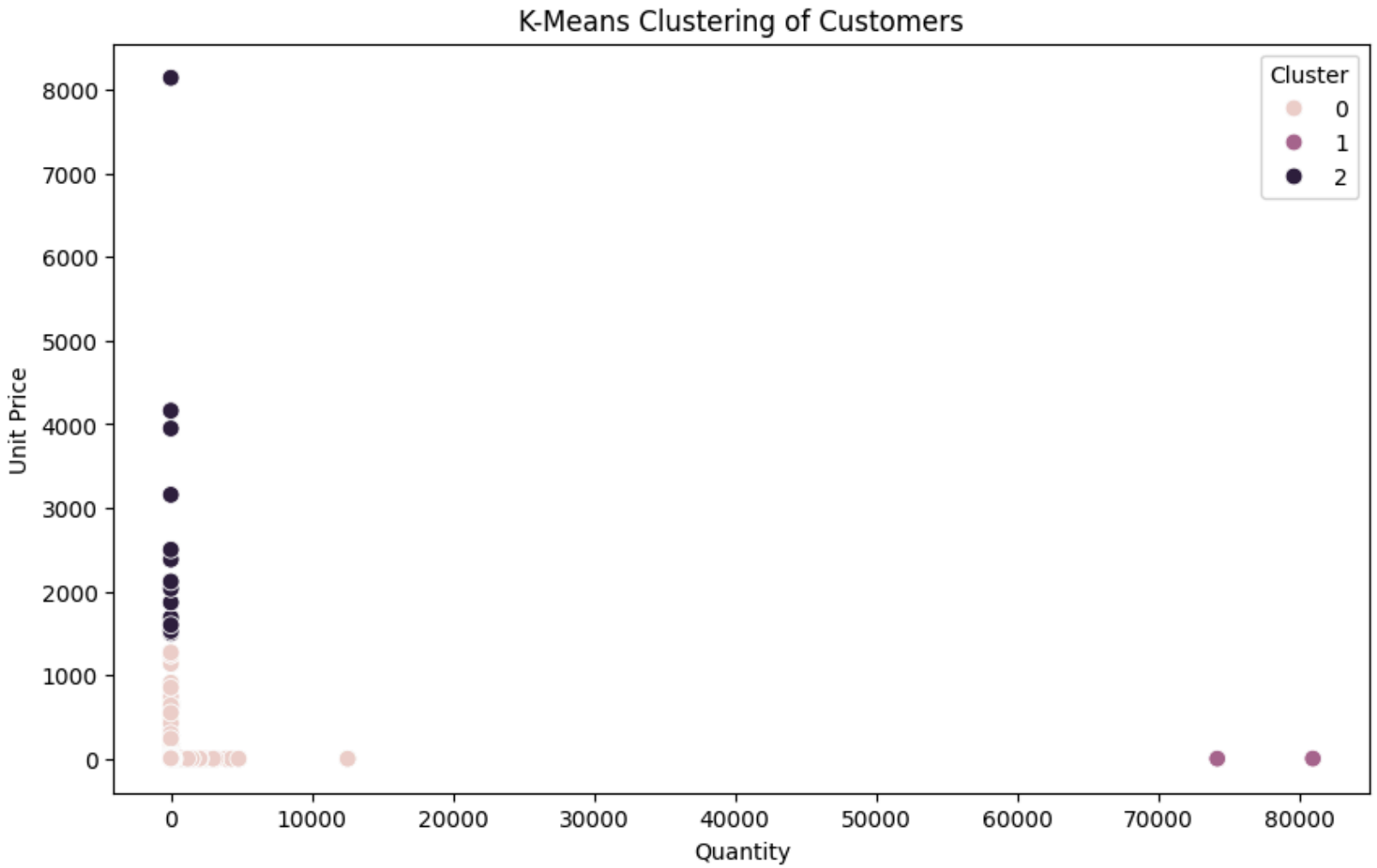
- K-Means Clustering was chosen for this project due to its simplicity, efficiency, and effectiveness in partitioning large datasets into distinct customer segments. As a widely used clustering algorithm, K-Means excels in scenarios where the goal is to categorize data into a predefined number of clusters, making it ideal for identifying distinct groups of customers based on purchasing behavior. Its iterative refinement process ensures convergence to optimal centroids, allowing for clear and interpretable segmentation. Additionally, K-Means is computationally efficient, making it suitable for handling large datasets like "Online Retail.csv," providing valuable insights for targeted marketing strategies and business decision-making.



- This approach is predicated on the assumption that k=3 is a local optimmum, while k=5 should be selected as the number of clusters. This method is deemed to be superior as it renders the determination of the optimal number of clusters more critical and transparent. However, it should be noted that this calculation is computationally intensive, as the coefficient must be computed for each case

Visualize the Clusters

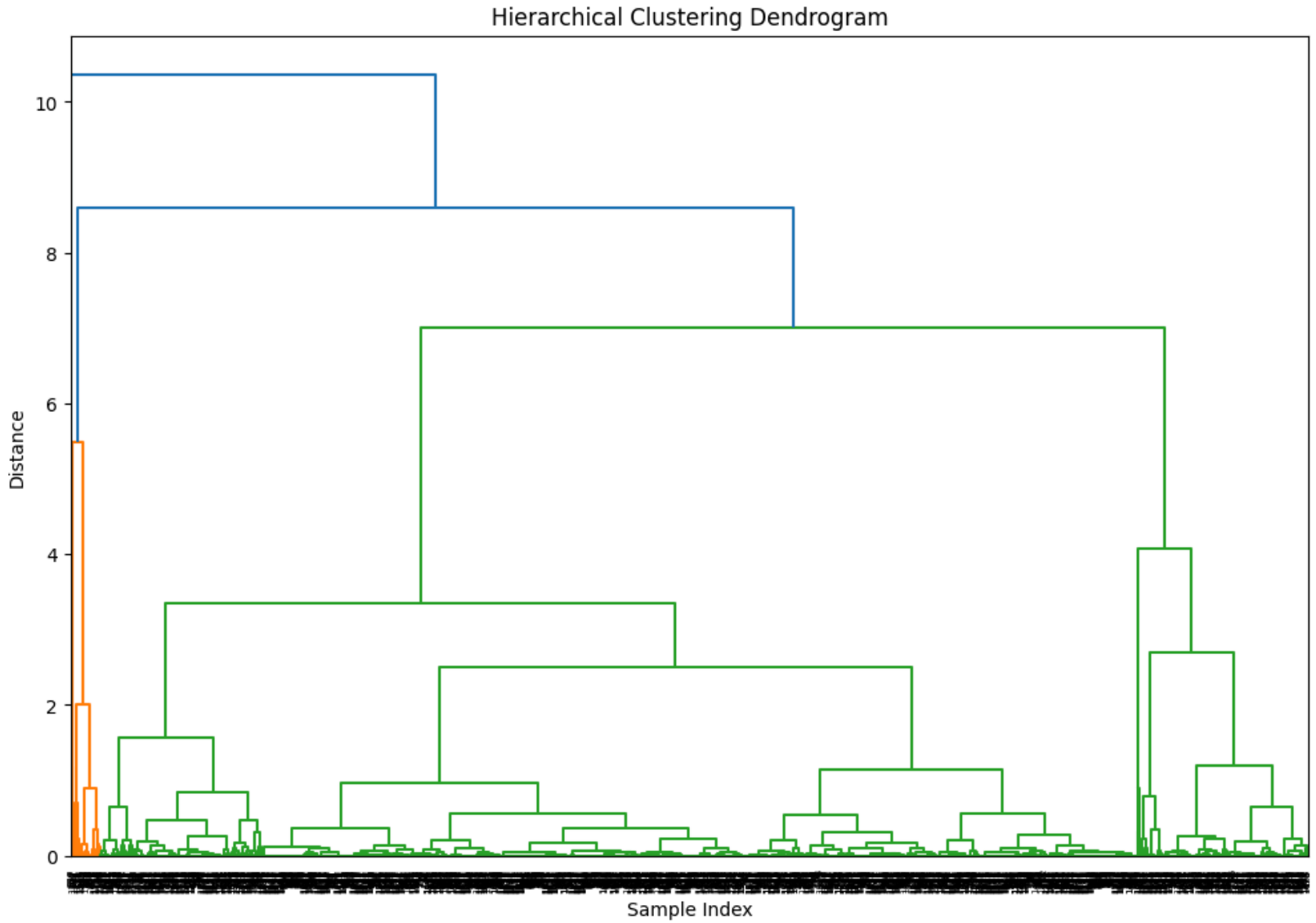
The clusters were visualized to understand customer segments based on their purchasing behavior.



Model 2: Hierarchical Clustering

- Hierarchical clustering is a powerful unsupervised learning technique that is well-suited for identifying meaningful clusters in customer data. Unlike K-Means, hierarchical clustering doesn't require specifying the number of clusters in advance.
- Using the 'ward' linkage method, we computed the hierarchical clustering dendrogram. This dendrogram was analyzed to determine the optimal number of clusters by visually inspecting where the largest vertical distances occur.

- Sample data based on 2000 rows from the pre-processed dataset.



Insights:

- The hierarchical clustering revealed key customer segments, which can be targeted with tailored marketing strategies. For instance, customers in high quantity but low unit price segments might be targeted with upselling campaigns, while high unit price segments could be offered premium services.

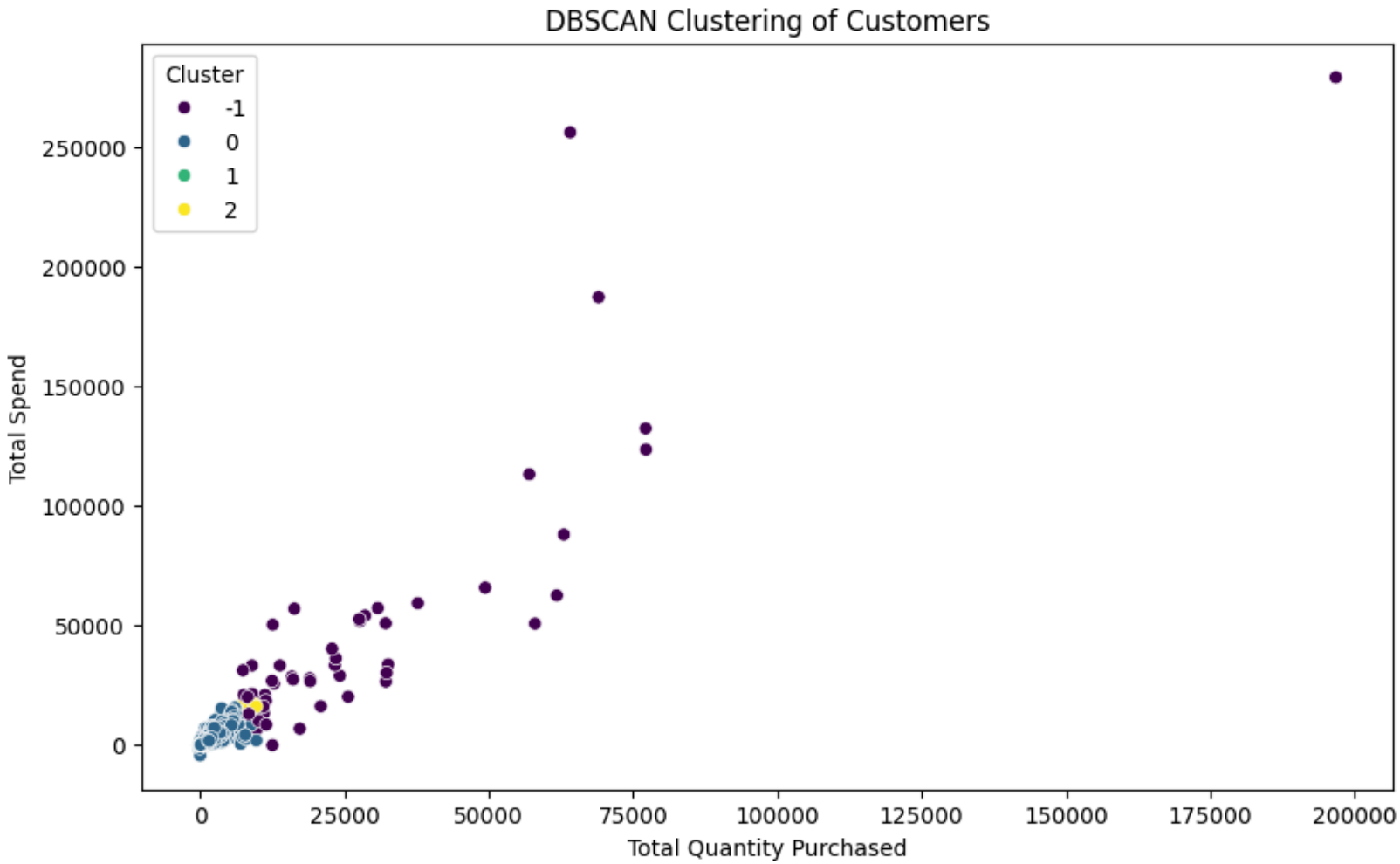
Model 3: DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is an excellent choice for customer segmentation because it effectively handles noise and identifies clusters of arbitrary shapes. Unlike K-Means, which assumes spherical clusters and requires the number of clusters to be specified in advance, DBSCAN forms clusters based on the density of data points. This allows for more natural groupings of customers, accommodating complex purchasing patterns without being distorted by outliers.

Moreover, DBSCAN's parameters, eps (maximum distance between two points to be considered neighbors) and min_samples (minimum number of points to form a dense region), adapt to the data's inherent structure. This flexibility makes DBSCAN robust in dealing with noise and discovering meaningful segments without prior knowledge of the number of clusters. These features set DBSCAN apart from K-Means and hierarchical clustering, making it particularly well-suited for customer segmentation tasks where data can be noisy and clusters are not necessarily spherical or uniformly sized.

Visualization of DBSCAN Clustering Results

The results are plotted using matplotlib to visualize the clusters. Noise points are marked in black.



5. Recommended Model

After evaluating K-Means Clustering, Hierarchical Clustering, and DBSCAN for customer segmentation on the "Online Retail.csv" dataset, DBSCAN emerges as the best and most recommended model. The nature of the retail dataset, which often includes a mix of dense clusters of customer behaviors and sparse outliers, makes DBSCAN particularly suitable. Unlike K-Means, which struggles with non-spherical clusters and requires the number of clusters to be predefined, DBSCAN can identify clusters of varying shapes and sizes based on the density of data points, and it inherently handles noise and outliers.

Additionally, DBSCAN does not necessitate specifying the number of clusters in advance, which can be advantageous given the potentially unknown number of customer segments in the retail data. Hierarchical Clustering, while useful for understanding the nested relationships between data points, can be computationally intensive and less effective in identifying the natural clusters in large datasets like this. DBSCAN's ability to form clusters based on local density variations and its robustness against outliers make it the most suitable choice for uncovering meaningful customer segments in the "Online Retail.csv" dataset, leading to more actionable insights for the business.

6. Key Findings

1. **Customer Segments:**
    - **K-Means:** Identified distinct customer segments based on purchase behavior but struggled with outliers.
    - **Hierarchical Clustering:** Provided a detailed hierarchy of customer segments, useful for understanding nested relationships but less scalable.
    - **DBSCAN:** Successfully identified clusters with varying densities and managed outliers effectively, showcasing its robustness in retail data.
2. **Outlier Detection:**
    - **K-Means and Hierarchical Clustering:** Limited in handling outliers, leading to potential misclassification or skewed results.
    - **DBSCAN:** Effectively detected and classified outliers as noise, providing cleaner and more accurate segmentation.
3. **Cluster Interpretation:**
    - **K-Means:** Easy to interpret clusters but required manual tuning of the number of clusters.
    - **Hierarchical Clustering:** Visual dendrogram offered insights into the data structure, but practical application on large datasets was challenging.
    - **DBSCAN:** Automatically determined the number of clusters, facilitating easier interpretation and application.

Insights

1. **Marketing Strategies:**
    - By identifying distinct customer segments, the business can tailor marketing campaigns to target specific groups more effectively, increasing engagement and conversion rates.

- DBSCAN's ability to manage outliers ensures that unusual customer behaviors do not skew the segmentation, allowing for more accurate targeting.

2. **Customer Retention:**

   - Understanding the characteristics of different customer segments enables the business to develop personalized retention strategies, reducing churn and enhancing customer loyalty.
   - Insights from hierarchical clustering can inform long-term retention plans by revealing nested relationships and hierarchical structures within the customer base.

3. **Operational Efficiency:**

   - Segmentation helps in optimizing inventory management by predicting the demand patterns of different customer groups.
   - With DBSCAN, the business can identify high-value customers and prioritize resources to cater to their needs, driving overall operational efficiency and profitability.

7. Suggestions for Next Steps

- To further enhance the insights and effectiveness of customer segmentation, it is recommended to integrate additional data sources, such as demographic information, customer feedback, and web analytics, to enrich the analysis. Exploring advanced clustering techniques like Gaussian Mixture Models (GMM) and incorporating dimensionality reduction methods such as PCA can help in handling high-dimensional data more effectively. Additionally, conducting a time-series analysis of customer purchase behavior could provide deeper insights into seasonal trends and changing patterns over time. Finally, validating the model with real-world business outcomes and continuously refining the segmentation approach based on ongoing data collection and feedback will ensure that the segmentation remains relevant and actionable for the business.